November 6, 2024

E-Filed

The Honorable Thomas S. Hixson
United States District Court for the Northern District of California
San Francisco Courthouse, Courtroom E – 15th Floor
450 Golden Gate Avenue
San Francisco, CA 94102

Re: *Kadrey, et al v. Meta Platforms, Inc.*; Case No. 3:23-cv-03417-VC

Dear Magistrate Judge Hixson:

Plaintiffs in the above-captioned action ("Plaintiffs") and Defendant Meta Platforms, Inc. ("Meta") jointly submit this omnibus letter brief regarding Plaintiff's outstanding disputes for "existing written discovery." *See* Dkt. No. 253. This brief is filed along with a stipulation asking for the Court's approval to combine Plaintiffs' five separate discovery motions ("Issues 1-5") into one omnibus joint letter. As set forth below, Plaintiffs present their argument on each Issue, followed by Meta's responsive position on each such issue.

/s/ Maxwell V. Pritt
Maxwell V. Pritt
Boies Schiller Flexner, LLP
Attorneys for Plaintiffs

/s/ Bobby Ghajar

Bobby Ghajar
Kathleen Hartnett
Phillip Morton
Cooley LLP
Attorneys for Meta

ISSUE #1: SEARCH TERMS, DATA SOURCES, AND PRODUCTION SIZE

Plaintiffs' Position

Meta has produced only 20,000 or so documents to date, excluding training data, a number that would have raised the specter of incompleteness and cherry picking even 15 years ago, to say nothing of 2024 discovery about a commercially-focused AI department with hundreds of team members. After significant follow-up, Plaintiffs learned for the first time just three weeks ago that Meta has excluded many (if not most) centralized, relevant sources of information as off-limits. In Meta's view, standard, central systems—paradigmatic *non*-custodial sources—should be treated as custodial and thus, to the extent used by any Meta employees other than the identified custodians, simply left out. Specifically, during an October 16, 2024 meet-and-confer with Meta regarding discovery deficiency letters (that Plaintiffs served on October 9), Meta acknowledged that it only searched and produced work email and other non-custodial company data sources like Workplace for its ten designated, self-selected document custodians—*no other employees*.[1] Other sources, such as WhatsApp, appear to have never been searched in a systematic manner at all. Meta's searches in response to Plaintiffs' RFPs are deficient and must be supplemented.

I. The Deficiencies With Meta's Searching are Evidenced by Low Production Volumes for Its Limited Custodians

Of the 10 custodians it originally self-selected, Meta produced thousands of documents from some custodians, but very few documents from others despite their heavy involvement in the technologies central to this action. For example, Meta produced 194 documents from Mike Clark despite the fact that he is a Director of Product for Generative AI at Meta, was involved in Meta's anti-scraping efforts (i.e., Meta's rules and policies surrounding what is (and what is not) permissible for Meta to scrape from the web), and that Meta has deemed him sufficiently important as to be the corporate designee for a plurality of the 30(b)(6) topics. Chaya Nayak Dep. Tr. at 187:5-188:15. Similarly, Meta produced only 335 documents for custodian Joelle Pineau (the vice-president of AI Research at Meta), and only 274 documents from the custodial file of Chris Marra (another Director of Product at Meta). Plaintiffs' bases for believing there are more relevant documents, in addition to reasonable inferences from the custodians' job functions, is the discrepancy between these numbers and other custodians such as Aurelien Rodriguez, for whom 9,503 documents have been produced.

As required under the ESI Order, on September 19, 2024, Meta provided Plaintiffs with a list of sources searched. Ex. A. That list, however, reveals glaring deficiencies. An obvious example is WhatsApp. While Meta now alleges that it searched WhatsApp for the 15 custodians and found a few hits, Dkt. No. 247 at 37, its September 19 list omits WhatsApp entirely. Plaintiffs first received six WhatsApp messages for a single custodian, Ahmad Al-Dahle, which were produced the night before his deposition. Plaintiffs then received another six WhatsApp messages for Joelle Pineau on November 4, approximately 36 hours before her deposition. There has been no explanation of why these WhatsApp messages first surfaced right before each custodian's depositions. Another example is text messages. Meta's witnesses have testified that they possess company-issued cell phones that they use for work-related communications. Melanie Kambadur Dep. Tr., 334:22-335:2. It appears these sources were never searched—a conclusion confirmed

[1] Meta identified two RFPs for which it did not limit its search to its ten custodians' email but refused to identify whether there were any additional RFPs for which it did not similarly limit its searches.

by Meta's witnesses in deposition, who have testified that their phones and computers were never imaged or taken by Meta for discovery in this case.

II. The Court Should Order Meta to Identify and Search All Relevant Non-Custodial Sources

Meta's side-stepping of relevant document sources is inadequate for a large, sophisticated entity. Meta knows that far more of its employees possess responsive information, and it cannot ignore all of those documents merely by limiting its search of central repositories with known sources of relevant information to a handful of custodians.[2] In addition to searching the files of the identified custodians, Meta has an affirmative duty under the ESI Order to identify and produce other known responsive documents that are housed by other employees or that are non-custodial in nature. Dkt. No. 101 at 6 ("Specific, non-duplicative ESI that is identified by a party as responsive to a discovery request shall not be withheld from review or production solely on the grounds that it was not identified by . . . the protocols described in, or developed in accordance with, this Order"); *see also In re Facebook, Inc. Consumer Privacy User Profile Litig.*, 2021 WL 10282215, at *2 (N.D. Cal. Sept. 29, 2021) ("The ESI Protocol agreed to by the parties does not limit Facebook's discovery obligations to only the identified custodians."). Meta's non-custodial sources almost certainly contain a substantial volume of responsive documents that were never searched, collected, or produced. Meta apparently searched these non-custodial sources for communications *involving just its ten (and now 15) document custodians*, Dkt. No. 247 at 37, but ignored the fact that hundreds of other employees invariably sent emails, communicated using Workplace, saved files to Workplace, and sent WhatsApp messages that are responsive to many of Plaintiffs' RFPs. p

III. The Court Should Also Require Prompt Disclosure of Search Terms for Meta's Custodians

As related to both previous sections, Meta has not yet divulged the (likely narrow) search terms that it used to search the files of any of its document custodians. Disclosure of search terms, in addition to being required by the ESI Order, is a fundamental feature of discovery that *usually* occurs voluntarily. *United Ass'n of Journeyman & Apprentices of the Plumbing & Pipe Fitting Indus., Underground Util./landscape Loc. Union No. 355 v. Maniglia Landscape, Inc.*, 2019 WL

---

[2] If the Court allows Meta to effectively leave out non-custodial sources, then Plaintiffs respectfully request that the Court order Meta to add the following 12 custodians: Brian Gamido (assigned task of reaching out to content owners to discuss potential licensing opportunities); Elisa Garcia Anzano (assigned task of reaching out to content owners to discuss potential licensing opportunities); Xavier Martinet (research engineer at FAIR with knowledge of books in Meta's training datasets); Guillame Lample (former research scientist at FAIR with knowledge of books in Meta's training datasets); Chris Cox (knowledge of and decisionmaker on licensing and LLM priorities); Moya Chen (research engineer at FAIR with knowledge of downloading and processing of LibGen training datasets); Susan Zhang (knowledge of Meta's downloading and use of LibGen); Sy Choudhury (lead role in licensing strategy); Jacob Xu (stated Meta's took Books3 "for free"); Arun Rao (key LLM product manager responsible for integrating Llama into Meta's full suite of commercial products); Eugene Nho (data acquisition for GenAI training); and Thomas Scialom (knowledge of Meta's fear of playing catch up and "clean up" of "problematic data"). 27 total custodians is minimal compared to other large Meta litigations. *See In re Facebook, Inc. Consumer Privacy User Profile Litig.*, 2021 WL 10282213, at *12 (N.D. Cal. Nov. 14, 2021) (Meta submission to special master stating it produced over 500,000 documents from 81 custodians, plus non-custodial sources).

7877821, at *2 (N.D. Cal. July 25, 2019) ("Parties exchange this sort of information as part of the general discovery meet-and-confer process"); *De Abadia-Peixoto v. U.S. Dep't of Homeland Sec.*, 2013 WL 4511925, at *4 (N.D. Cal. Aug. 23, 2013) (ordering defendant to disclose search parameters and to meet and confer regarding their sufficiency).

IV. Prayer for Relief

Plaintiffs respectfully request that: (a) the Court order that non-custodial sources of relevant information should be subject to reasonable and proportional searching (or alternatively, that Meta expand its list of custodians) and (b) that Meta disclose its search terms. Plaintiffs request that the Court (1) order that the parties exchange the search terms run and data sources searched for all requests and all custodians, the resulting hit counts, and any non-custodial data sources that were not searched but may contain relevant information, on **November 13**; (2) require the parties to meet and confer by **November 15** to discuss whether any additional search terms need to be run or data sources searched; and (3) set a **November 18** deadline for a joint letter brief on any unresolved issues involving search terms and data sources.

V. Plaintiffs' Dispute Is Timely

Finally, Plaintiffs' dispute about Meta's continuing search and production efforts continues to be ripe, and there remains sufficient time during ongoing fact discovery for the Court to grant meaningful relief. This motion is not an attempt to "redo all of Meta's document productions[.]" Dkt. No. 252 at 2.[3] Meta is in the process of responding to new RFPs issued by Plaintiffs in October, and it presumably will employ similar search methodologies to those used before. Further, parties have an ongoing duty to supplement their responses to RFPs if they learn their prior productions were incomplete in some material respect. Fed. R. Civ. P. 26(e)(1). It is not uncommon for courts to direct parties to search new data sources near the end of fact discovery if those data sources should have been searched much sooner. *Owen v. Hyundai Motor Am.*, 344 F.R.D. 531, 535 (E.D. Cal. 2023) (compelling defendant to run new search terms and identify responsive data sources in dispute raised at end of fact discovery period).

Defendants' Position

Plaintiffs' motion loses sight of the lone copyright claim at issue in this case, what the parties need to brief summary judgment on fair use, and the full year of discovery already conducted. Just 35 days remain in fact discovery, and despite the Court's admonitions last week that "[w]e're not going to redo all of Meta's document productions at this point" and that "[i]f Plaintiffs had a problem with Meta's search terms, date ranges, metadata collection, and so on, those issues should have been raised a long time ago," ECF No. 252 at 2, Plaintiffs in "Issue #1" seek to entirely redo Meta's document production and to rewrite the ESI Order long ago agreed to by the Parties and entered by the Court that has governed discovery throughout this case. Plaintiffs' belated complaints are not only untimely and meritless, but it would eviscerate Rule 26 and the negotiated, Court-approved limitations on custodial searches for ESI (which the parties agreed would be limited to 10 custodians per side selected by the producing party), ECF No.

[3] Even for Meta's 10 initial custodians, Plaintiffs have raised with Meta the problems with its "search terms, date ranges, metadata collection, and so on." Dkt. No. 252 at 2. Meta should not be permitted to get away with its failure to comply with basic discovery obligations in a landmark case just because those issues could have been raised earlier in discovery, before Plaintiffs reconstituted their legal team at Judge Chhabria's direction.

101 at 8, and which was expanded to 15 Meta custodians by the Court one month ago, ECF No. 212). Not only is there no need for the massive expansion of discovery that Plaintiffs now seek, but if Plaintiffs' requested relief is granted, it would result in a wholesale redo of Meta's document production, requiring search, collection, review, and production of ESI from untold number of custodians, far in excess of anything contemplated throughout this case and vastly disproportionate to the needs of the case. Such a dramatic expansion of discovery at this late stage of the case would make completing discovery by December 13 impossible. For all these reasons, Plaintiffs' motion regarding "Issue #1" should be denied

Plaintiffs premise their demand for a massive expansion of discovery on the unsubstantiated speculation that Meta's document production is not "big" enough, assuming—incorrectly—that there must be more documents addressing the narrow set of issues at play in this case.[4] Tellingly, Plaintiffs do not identify a single RFP with a deficient response or any category of documents allegedly missing in Meta's production. To be clear, Meta's document productions to date, both custodial and non-custodial, were based on the RFPs served by Plaintiffs. Meta has produced the documents it said it would produce in response to those RFPs, and it was not until October 9th that Plaintiffs raised for the first time any issue with Meta's responses to those RFPs. As Plaintiffs also acknowledge, Meta will be making additional productions in response to Plaintiffs' 5th and 6th sets of RFPs (served on Oct. 9 and Oct. 18) and for the 5 custodians newly added per Court order. Meta anticipates producing thousands more custodial and non-custodial documents in response to those document requests, including several thousand being produced **today**.

Not only is Plaintiffs' motion premised on speculation rather than evidence of any alleged failure by Meta, but the ESI Order does not support and in fact contradicts Plaintiffs' demands. The only provision of the ESI Order cited by Plaintiffs is paragraph 7.d, which provides that "[s]pecific, non-duplicate ESI that is identified by a party as responsive to a discovery request shall not be withheld from review or production solely on the grounds that it was not identified by (or is subject to an exclusion set forth in) the [ESI Order protocols." ECF No. 101 at 12. But this provision is a narrow exception to the ESI order's custodial search and production protocols. Indeed, paragraph 7.d simply requires the parties to review and produce *specific*, non-duplicative and responsive ESI that happens to be identified outside of the ESI order's custodial search and production protocols. Nothing in paragraph 7.d obligates a party to conduct broad searches of custodial ESI (such as emails and chats) across untold numbers of people—effectively eviscerating the ESI Order's clear directive that ESI discovery is governed by the specific custodial search and production protocols that Meta has been following. Plaintiffs' interpretation of this narrow exception in the ESI Order would render the entire meet and confer process mandated by Federal Rule 26(f) as implemented by the ESI protocol meaningless.

With respect to WhatsApp messages in particular, Plaintiffs' complaints are similarly unfounded. Meta inquired about its custodians' use of various platforms, including WhatsApp,

[4] Meta has also produced over 15 terabytes of data, and made multiple gigabytes of source code available for inspection. Moreover, Plaintiffs also miscount the number of documents associated with Ms. Pineau, who has 355 custodial documents produced. And Chris Marra was *not an identified custodian* (a fact Plaintiffs fail to acknowledge), and the inclusion of his documents demonstrates that Meta in fact searched beyond its 10 original custodians for responsive documents.

for work-related communications, and when a custodian identified WhatsApp as such a platform, Meta conducted searches of that application. Only a handful of the 15 ESI custodians (e.g., Mr. Al-Dahle, Ms. Pineau) used WhatsApp for a limited number of work-related communications, and Meta collected and produced those responsive WhatsApp messages. Yet Plaintiffs are demanding that Meta conduct searches of employees' personal WhatsApp messages for employees that have confirmed they do not use WhatsApp for work (e.g., Chaya Nayak and Sergey Edunov, who testified under oath they do not use WhatsApp for work). Plaintiffs cite no authority for such an unnecessary intrusion into employees' personal, non-work related communications.[5] *See Int'l Longshore & Warehouse Union v. ICTSI*, 2018 WL 6305665 *3 (D. Ore. Dec. 3, 2018) (denying motion to compel personal email account information from custodians because the moving party did not show more than a de minimis use of personal email for business) (citing *Matthew Enter., Inc. v. Chrysler Grp. LLC*, 2015 WL 8482256, at *3 (N.D. Cal. Dec. 10, 2015) (denying motion to compel production of employees' personal email accounts and noting that the moving party had not identified any legal authority under which the employer could force its employees to turn over email from personal accounts)).

Implicitly recognizing that the ESI Order does not support their request for a massive expansion of custodial discovery, Plaintiffs seek to frame their request as seeking "non-custodial" or "centralized" data. To be clear, Plaintiffs are seeking broad-based searches of data sources that would be the subject of a custodial ESI search under the ESI Order, specifically email and Workplace chat messages. Moreover, Meta has already disclosed and addressed the "additional data sources with potentially responsive information" under the ESI order. ECF No. 101 at 6(c). As identified in Meta's September 19, 2024 disclosure of data sources, Meta has already reviewed and produced from a comprehensive list of sources including E-mail, Workplace Chat, Contracts, Google Suite, Github, S3 (AWS servers), Workplace Groups[6] and others. These include sources the Plaintiff might also classify as "non-custodial" or "centralized" data sources, such as relevant public websites, Google Share drives (part of the Google Suite) and Github. Consistent with the parties agreement in the ESI Order, the Producing Party (Meta) was best situated to determine the most appropriate method(s) to search, collect, cull and produce discovery from all of the data sources Meta identified. ECF No. 101 at paragraph 7.

---

[5] Plaintiffs also complain that Meta has not "taken" the phones or computers of employees for "imaging." This misunderstands how employees work and store information at Meta and also misrepresents Meta's discovery obligations. More specifically, there is no need to image every custodian's phone or computer in this matter because relevant sources are accessible through other methods. For example, Meta employees' Workplace chat messages (whether sent by phone or computer) are archived into Proofpoint, where Meta can collect those messages directly. Also, Meta's custodial interviews inquired about custodians' use of other platforms (if any) to conduct their work, and if a custodian had identified potentially responsive information uniquely stored on their phone or computer, Meta would have collected it.

[6] Workplace is a communications platform that securely combines chat, video, groups, and users' intranet. It is generally used to collaborate and share knowledge through notes and posts. Workplace contains the functions Groups and Chat where Groups allow users to comment and create posts while Chat is used as the internal instant messaging system. Workplace is internally developed and its data is stored on premises in Meta's internal databases.

Plaintiffs are instead asking the Court to order Meta five weeks before the close of fact discovery to extensively search additional *custodial* data sources like emails and Workplace chat communications of *non-custodians*, imposing massive ESI burdens far beyond those agreed to by the ESI Order, as previously expanded by the Court's recent order increasing the number of custodians from 10 to 15 per side. Plaintiffs have no support for this approach, which conflicts with the ESI Order. *In re Facebook, Inc. Consumer Privacy User Profile Litig.* does not support the relief they seek. There, the Court simply determined that an ESI Order with different language than the ESI Order entered in this case did not prevent production of physical notebooks. *In re Facebook, Inc. Consumer Privacy User Profile Litig.*, 2021 WL 10282215, at *2 (N.D. Cal. Sept. 29, 2021). That authority has no bearing here.

Plaintiffs' alternative request for 12 additional custodians (*supra* FN2) should also be rejected. The time for seeking new custodians passed long ago. If Plaintiffs needed more than the 5 extra custodians the Court already allowed, they should have asked for them when they briefed the issue 5 weeks ago. ECF No. 190 at 2. Given the burdens and time necessary to collect, search, review and produce custodial documents, those efforts will run long past the Court's fact discovery deadline. *See* ECF No. 190-8 at 2 (explaining time involved in custodial collection and production efforts); ECF 196 at 2 (recognizing that "[a]dding five document custodians is a big change, not a small change, to the scope of Meta's document production obligations in this case").

Finally, with respect to search terms, even though the Court has already said that ""[i]f Plaintiffs had a problem with Meta's search terms . . . [it] should have been raised a long time ago," to moot this issue Meta is prepared to mutually exchange its search terms and hit counts with Plaintiff next week and will work with Plaintiff to finalize the timing of that exchange. As Meta has already told Plaintiffs, there is no dispute about the exchange of such terms.[7] The Court should deny the requested relief as to Issue #1.

Plaintiffs' Reply

Meta's response does not address the key issue: Meta searched non-custodial databases for their custodians only. Data sources like company email servers, Meta-owned WhatsApp, and company tool Workplace, are not custodial data sources when the documents are centrally archived rather than housed on employees' devices. Meta effectively says as much, noting that there was no need to image its custodians' devices because all the same data was available through non-custodial sources, including a repository of centrally archived Workplace Chat messages for all employees. Defs' n.5; *In re Facebook, Inc. Consumer Priv. User Profile Litig*., 2021 WL 10282172, at *4 (N.D. Cal. Oct. 11, 2021) (noting a wide variety of non-custodial sources used at Facebook). If Meta had such ready access to this data for its employees, it should have run its search terms over the full data sources (or at least sources pertaining to the AI Department), not just for the 15 custodians.

This is why Meta's document productions are so small. Meta may be producing a few thousand more documents today, but even 25,000 documents is miniscule in a case involving a

[7] Among other unnecessary and impractical deadlines, Plaintiffs ask the Court to order additional letter briefing on search terms and data sources on November 18th. Such briefing is unnecessary and untimely under the Court's prior admonition that issues with Meta's search terms "should have been raised a long time ago." Moreover, Meta long ago disclosed its data sources as discussed above.

thousand-person AI department that has been working on AI and LLMs for many years. These problems are ongoing: Meta apparently will continue to search non-custodial sources solely for its 15 custodians in response to Plaintiffs' October RFPs absent a Court order to just run the same search terms more broadly. Further, Meta argues Plaintiffs never identified any specific RFPs receiving a deficient response. The problem is that Meta did not identify which produced documents respond to which RFPs.

Plaintiffs are glad Meta now states its willingness to exchange search terms. That position is in stark contrast to the silent approach taken by Meta over the past month in response to Plaintiffs' repeated requests for this exchange to take place. Because there is now no dispute that the parties should exchange search terms and hit counts, Plaintiffs ask the Court to order the exchange take place and set a deadline for the parties to raise any disputes regarding those search terms.

Defendant's Response to Plaintiffs' Reply

Plaintiffs claim that Meta "searched non-custodial databases for their custodians only," but this makes no sense. "Non-custodial databases" are centralized repositories of information not created or controlled by any one custodian, so it doesn't make sense to say that Meta conducted the searches of these databases for the custodians only. To the contrary, Meta collected documents from a wide variety of locations that are non-custodial, and spoke with many individuals in addition to its custodians to try to locate responsive documents. This is demonstrated by Meta's collection of Workplace posts, source code and datasets - data from quintessential "non-custodial" sources that span *terabytes* of data. Plaintiffs' complaint lacks any basis as they are unable to point to a single deficient response to Plaintiffs' RFP, focusing instead on imagined deficiencies based on the number of documents. Plaintiffs do not even attempt to justify their evisceration of the ESI order's carefully negotiated scope for the search, collection and production of ESI from custodians. Plaintiffs' reply makes clear that they are asking the Court to order Meta to search email and chat messages for at least 1000 custodians. Plaintiffs provide no justification for such an extreme expansion of discovery in this matter at such a late date.

As to the exchange of search terms, Meta has long told Plaintiffs that it was prepared to do a mutual exchange. Indeed, as recently as November 4, counsel wrote Plaintiffs "We remain, as previously stated, willing to agree to a mutually agreeable date for the exchange of search terms and hit counts … only one plaintiff has provided search terms (without a hit report). Please let us know when **all** Plaintiffs will be ready … We are prepared to do the same." Plaintiffs sent a follow up, but with the instant flurry of motions today, the parties have not yet agreed upon an exchange date. Meta remains ready and willing to provide this information. However, there is no basis to set a deadline to "dispute" those search terms this late in discovery, as any concerns regarding the size and scope of Meta's production *in response to RFPs served more than six months ago* should have been raised long ago. This is particularly the case here, given the timeframe left in discovery and Meta's continuing review of five new custodians and additional documents from custodial and non-custodial sources in response to Plaintiffs' more recent sets which no dispute is yet ripe. for

ISSUE #2: DISCOVERY RELATED TO LLAMA 4 PROGRAM

Plaintiffs' Position

Since early 2023, Meta has released new large language models ("LLMs") at frequent intervals. Meta's original LLaMA model ("Llama 1") was released on February 24, 2023, followed by Llama 2 on July 18, 2023; Llama 3 on April 18, 2024; Llama 3.1 on July 23, 2024; and Llama 3.2 on September 25, 2024. Meta is currently in the process of developing Llama 4 (the subject of this motion) and is scheduled to release this next iteration sometime in early 2025.

Throughout discovery in this case, Meta implicitly has conceded that successor models are relevant to Plaintiffs' claims even though Llama 1 was the operative version at the time Plaintiffs filed their original complaint,[8] and recently agreed that its corporate designee witnesses will also address Llama 4. This all makes sense: Models build on each other, and Meta's work on its previous models has informed its Llama 4 data strategies. Llama 4 is discussed in many of Meta's discovery materials, including key documents in this case. This evidence already makes clear that Meta's development of Llama 4—motivated by its desire to launch a best-in-class LLM that could beat OpenAI's GPT-4 and advance Meta's entire commercial portfolio—has continued to involve copyright infringement. Yet, Meta refuses to run searches for responsive documents involving Llama 4 and has not made available for inspection a complete set of source code and related data for all phases of its LLM training and implementation, including for Llama 4. Put simply, Llama 4 is relevant to this case under Rule 26 and is within the scope of permissible discovery.

I. Llama 4 Is Relevant to Plaintiffs' Allegations of Infringement.

Parties are entitled to discover any nonprivileged information that is reasonably calculated to lead to the discovery of admissible evidence. Fed R. Civ. P. 26(b)(1). The relevance standard under Rule 26 is broad. *Shoen v. Shoen*, 5 F.3d 1289, 1292 (9th Cir. 1993).

A. Llama 4 Infringes Plaintiffs' Copyrights Through the Same Mechanisms as Prior Llama Versions.

Llama 4 and its related documents are plainly relevant and discoverable under Rule 26. For one, Meta's use of "shadow libraries"—or online repositories comprising mass pirated copyrighted works—has continued and if anything increased over time. This evidence bears directly on Plaintiffs' claims in a few ways. First, it demonstrates ongoing willful infringement. Second, it shows that books datasets have been persistently significant because they comprise "high quality" data. This evidence also makes clear the commercial nature of Meta's investments in Generative AI through its Llama models, with each model helping Meta get closer to realizing its commercial aspirations.

---

[8] Plaintiffs defined the proposed class to encompass *all* versions of the Llama language models. Dkt. No. 133 at ¶ 83 (defining the proposed class in this case as "All persons or entities domiciled in the United States that own a United States copyright in any work that was used as training data for any version of the Llama language models between July 7, 2020 and the present."); *see also Farnsworth v. Meta Platforms, Inc.*, No. 3:24-cv-06893, Dkt. No. 1, at ¶ 55 (N.D. Cal. Oct. 1, 2024) (defining class to encompass owners of copyrighted materials used in current training for even unreleased models, and referring to books that "were or are used by Meta in the process of LLM training, research, or development, including but not limited to the training and development of its Llama models").

To date, Meta has produced a sparing subset of documents that mention Llama 4. These documents—just sufficient enough to show Llama 4's relevance, but far from a complete set—span the 2023 to 2024 timeframe. [redacted] For Meta, each Llama model aspires to surpass its precursor in terms of size, data, and, most importantly, ability to be integrated within and strengthen Meta's entire commercial portfolio (*i.e.*, Facebook, WhatsApp, and Instagram). The documents convey a clear linear progression whereby every model is building on an earlier one.[9] [redacted] Put differently, Llama 4 is no different than the prior Llama models so far as Rule 26 is concerned—all are relevant and discoverable.

B. That Llama 4 Is Still in Development Is Immaterial.

Meta's sole argument against producing responsive documents involving Llama 4 is that the model is still under development and will be released only after fact discovery concludes. Similar arguments about "in-development" products have been rejected by other courts in infringement cases. *E.g.*, *Bryant v. Mattel, Inc.*, 2007 WL 5430888, at *3 (E.D. Cal. May 18, 2007) (holding that "drawings and designs for unreleased products are also relevant to Mattel's copyright infringement claim" in light of Mattel's allegations of "ongoing conduct of reproducing and creating derivative works," and rejecting argument that "Mattel cannot possibly have suffered a cognizable injury relating to . . . products that have not yet been released"); *Paice, LLC v. Hyundai Motor Co.*, 2014 WL 3819204, at *12 (D. Md. June 27, 2014) (granting discovery into defendants' "future non-commercialized products alleged to have actually undergone an infringing use"); *Bigband Networks, Inc. v. Imagine Comms, Inc.*, 2010 WL 2898288, at *1 (D. Del. July 20, 2010) (granting discovery into the source code of defendants' future products on similar grounds). Moreover, Meta continues to downplay or avoid the basic issue in the case: Meta infringed Plaintiffs' copyrights when it downloaded their books.

Meta's use of copyrighted, protected material for training data constitutes copyright infringement. In other words, Meta committed copyright infringement in the first iteration when it downloaded and copied Plaintiffs' works from various online shadow libraries to make datasets for training Meta's LLM models—a time period that necessarily occurs well before the release date of any model. [redacted] To the contrary, Llama 4 is part of the very DNA of Meta's ultimate commercial vision for all of its Llama models—and it is also,

[9] As further evidence of the rapid, iterative nature of Meta's updates, [redacted] For this reason, Plaintiffs' recent RFPs broadly encompass all versions of Meta's AI models developed or in development, including Llama 4 and Llama 5.

potentially, the most infringing model to date. [redacted]

II. Request for Relief

To the extent that Meta's original search terms did not encompass Llama 4, Plaintiffs request an opportunity to propose a set of discrete additional search terms involving Llama 4 that Meta can run for all fifteen document custodians. And, if Meta has withheld documents responsive to its initial search terms on the ground that the documents pertain to Llama 4, Plaintiffs request that Meta revisit those documents and make supplemental productions accordingly.

Defendants' Position

As an initial matter, Meta's unreleased LLMs are not at issue in the operative Complaint and Plaintiffs have no copyright claim against as-yet-completed or unreleased LLMs. Nonetheless, Meta is not withholding otherwise responsive documents from production on account of the fact that they pertain to the upcoming Llama 4 model. Meta's search terms in this case would have and did encompass documents about Llama 4. Meta did not limit its searches for documents to any particular version of Llama, and has produced hundreds of documents that reference Llama 4. Moreover, as Plaintiffs acknowledge, Meta's witnesses have answered deposition questions relating to Llama 4. There is no dispute for the Court to address on this issue.

Plaintiffs also claim that Meta has not made available source code and related data pertaining to Llama 4, but as explained in Meta's response to Issue #5 below, this is not properly part of this motion because Plaintiffs never propounded any Requests for Production of Documents (RFPs) seeking Meta source code as part of *existing* written discovery—as opposed to as part of Plaintiffs' more recent discovery requests served in October 2024. As further noted below regarding Issue #5, Meta's responses to Plaintiffs' RFPs seeking source code (the only RFPs seeking source code) are due **today**, and as such, the parties have not even begun any meet-and-confer process to assess whether the parties can reach agreement on an acceptable supplement to the existing source code production. Without waiver of its objections to those pending RFPs, Meta intends to provide make source code related to Llama 4 available for inspection (to the extent not already produced) in response to Plaintiffs' new RFPs and will continue to work with Plaintiffs on source code-related issues. Plaintiffs' attempt to seek premature adjudication on unripe source code issues through this motion should be rejected.

Plaintiffs' Reply

As has often been true in this litigation thus far, the position Meta reports to the Court bears no relation to the position Meta has taken with Plaintiffs. Meta objected to the inclusion of Llama 4 in its written responses and objections to ***all*** of Plaintiffs' discovery requests, including most recently today, November 8, in its responses to Plaintiffs' Fifth Set of RFPs, re-defining "Meta Language Models" so that it did ***not*** include all versions of Llama, including Llama 4 and Llama 5: "Meta construes "Llama Models" to mean the models within the Llama family of LLMs that have been publicly released by Meta, namely, Llama 1, Llama 2, Code Llama, and Llama 3." Yet, for first time in this brief, Meta simultaneously states that it drops its objections and claims its searches "would have and did encompass documents about Llama 4."

So, which of Meta's conflicting positions is to be believed? Given how few produced documents mention Llama 4 (or 5), Meta's carefully worded claim that its still-undisclosed "search terms … did encompass documents about Llama 4" is not credible. Indeed, what that closely-lawyered phrasing probably means is that a few search terms picked up a few documents that mention Llama 4; but Meta notably does *not* say it specifically searched for Llama 4 (and 5) documents in response to Plaintiffs' discovery requests that sought relevant information regarding all Llama models, including Llama 4 and 5. In any event, given that Meta now drops any objection to searching for these documents, the Court should enter the relief requested as it is now undisputed that Plaintiffs are entitled to Llama 4 and 5 documents. The same goes for Llama 4 source code—Meta now says it will make it available. So it is now also undisputed that Plaintiffs are entitled to have Meta make available the source code for all Llama models for all phases of the LLM process as it exists in the ordinary course at Meta, and the Court should enter the relief requested.

Defendant's Response to Plaintiffs' Reply

Plaintiffs articulate no proportional basis for demanding that Meta search for and produce all things relating to unreleased LLMs, including Llama 4. There is nothing about Meta's efforts with respect to Llama 4 that make any fact in dispute more or less likely. Notably, Plaintiffs' reply fails to point to any claim in their Complaint that would relate to a future LLM that is not released. They cite no case law establishing the relevance of future, in-development models to a claim of copyright infringement based on the training of completed and publicly released models. Moreover, any perceived lack of information is more reflective of the status of Llama 4's development than any perceived discovery limits that were not put in place by Meta. The bottom line, as Plaintiffs acknowledge, is that Llama 4 and potential future iterations of Llama models are not complete, such that a fishing expedition into highly sensitive competitive information relating to future LLM models is not relevant or proportionate to the needs of this case.

Separately, Meta's willingness to provide documents and code commensurate with the current status of training Llama 4 (and not any or all source code for unreleased models, regardless of the relevance of the code to the issues in this case) should and cannot be construed as a concession regarding the relevance or proportionality of Plaintiffs' demands for more. Thus, there is nothing for the Court to order and Plaintiffs' request should be denied.

ISSUE #3: RESPONSES TO PLAINTIFFS' INTERROGATORIES

Plaintiffs' Position

Plaintiffs have issued three sets of interrogatories to Meta—on December 27, 2023 (Set 1, Nos. 1-15), on August 29, 2024 (Set 2, Nos. 16-17), and on October 10, 2024 (Set 3, Nos. 19-48). ("Interrogatories" or "ROGs").[10] Ex. B. Meta initially refused to respond to Interrogatory Nos. 13-15, arguing that some contained subparts and thus counted towards the presumptive limit of 25 interrogatories. Ex. B. Meta subsequently agreed on October 21 to supplement its responses to Interrogatory Nos. 13 and 14 but maintained its refusal to answer Interrogatory No. 15, which seeks an explanation of the role of certain identified persons in developing and marketing Llama. Yet, Meta also answered Interrogatory Nos. 16 and 17 in Set 2 (on September 30), and Meta served its own set of interrogatories exceeding the presumptive limit eight days *after* Plaintiffs served their Set 3. Meta cannot unilaterally decide which interrogatories to answer and has no proper basis to withhold additional responses.

I. Meta Must Answer Interrogatory No. 15.

Meta's argument that it need not respond to Interrogatory No. 15 because of its subparts is misplaced. The rule regarding impermissible subparts in interrogatories applies only to discrete subparts that ask about separate and distinct subjects; subparts count as one interrogatory so long as they are logically or factually related. *Stamps.Com v. Endicia*, 2009 WL 2576371, at *3 (C.D. Cal. May 21, 2009) ("[C]ourts generally agree that 'interrogatory subparts are to be counted as one interrogatory . . . if they are logically or factually subsumed within and necessarily related to the primary question.") (*Trevino v. ACB Am., Inc.*, 232 F.R.D. 612, 614 (N.D. Cal. 2006); *Synopsys v. ATopTech*, 319 F.R.D. 293, 297 (N.D. Cal. 2016) ("Subparts asking for facts, documents, and witnesses relating to a primary contention or allegation are logically or factually related, and thus should be construed as subsumed in the primary question."). Each Set 1 interrogatory (Nos. 1-15) properly addresses a single, discrete (and relevant) topic.[11] For example, the disputed Interrogatory No. 15 asks for an explanation of the role that 15 enumerated Meta employees played in developing its LLMs. Under Defendants' view, that interrogatory alone would constitute 60 percent of the total interrogatories available to Plaintiffs under Rule 33. This cannot be right.

Regardless of whether subparts constitute separate interrogatories, Meta waived its objection to Interrogatory No. 15 when it later answered Plaintiffs' Interrogatory Nos. 16-17 in a subsequent

---

[10] Plaintiffs served a total of 47 interrogatories, having inadvertently skipped No. 18.

[11] Those topics are: (ROG 1) nature and source of data used to train Llama, (ROG 2) process and persons responsible for modifying Llama, (ROG 3) Reinforcement Learning with Human Feedback (RLHF) process for Llama, (ROG 4) pre-release assessment of risk, safety, and alignment of Llama, (ROG 5) agreements showing a financial interest of Meta insiders with respect to Llama training data, (ROG 6) identification of Meta's directors, officers, and board members, (ROG 7) employees responsible for development of Llama and ethical or legal concerns, (ROG 8) software, databases, or services used to develop Llama, (ROG 9) contact information for potential witnesses identified in ROGs, (ROG 10) identification of persons from whom Meta licensed or purchased training data, (ROG 11) identification of individuals/entities who own stock in Meta above 5%, (ROG 12) identification of individuals who planned or developed Llama, (ROG 13) identification of Llama training datasets and the policies followed to include or exclude datasets, (ROG 14) identification of individuals granted or denied access to Llama 1, and (ROG 15) an explanation of the role of identified persons in developing and marketing Llama.

set. Ex. B. Meta reiterated its objections to Interrogatory Nos. 16 and 17 as exceeding Rule 33's limit, but Meta then proceeded to answer them. If Meta wanted to preserve the numerosity objection, it needed to apply its objection uniformly. What Meta cannot do is respond to some of the implicated interrogatories out-of-order while refusing to respond to other, earlier ones. *See Capacchione v. Charlotte-Mecklenburg Schools*, 182 F.R.D. 486 492 n.4 (W.D. N.C. 1998) (explaining that the recourse for a party seeking to preserve "objections to supernumerary interrogatories is to answer up to the numerical limit and object to the remainder without answering.").

II. Meta Must Supplement its Answer to Interrogatory No. 17.

One of the nonconsecutive interrogatories that Meta *did* choose to answer—Interrogatory No. 17—involves a potential advice-of-counsel defense. Meta answered the question by stating it "does not ***presently*** intend to assert the advice of counsel defense." *Id*. Meta should be required to provide a definitive "yes" or "no" answer.

When a defendant asserts reliance on the advice of counsel as a defense, attorney-client privilege is waived over all communications with counsel related to that advice. *See*, *e.g.*, *SNK Corp. of Am. V. Atlus Dream Entertainment Co., Ltd.*, 188 F.R.D. 566, 571 (N.D. Cal. 1999) ("courts have found the injection of the advice of counsel to waive the attorney-client privilege as to communications and documents relating to the advice.") (cleaned up). With minimal time remaining in the fact discovery period, Plaintiffs need to definitively know now—not sometime in the future—whether Meta may assert this defense. Meta's attempt to leave the door open for an advice of counsel defense down the road is improper and is the type of tactic that could lead to "a train wreck at the close of fact discovery," Dkt. No. 231 at 3. Meta therefore needs to supplement its answer to Interrogatory No. 17 to state definitively whether it is asserting this defense.

III. Plaintiffs Request Leave to Issue Additional Interrogatories and an Order Requiring Meta to Answer Them.

Meta also refuses to respond to any of Plaintiffs' interrogatories in Set 3 (Nos. 18-43). First, this discovery was timely. Plaintiffs were diligent in serving Set 3 in response to Judge Chhabria's 60-day extension of discovery, ordered on October 4, 2024. Dkt. 211. Plaintiffs served Set 3 less than a week later, on October 10, over two months before the close of discovery, and Meta's deadline to respond is November 11. Judge Chhabria's 10-week extension of discovery on October 4 was for all purposes, without restrictions on either party seeking additional discovery or issuing new requests. Dkt. 211. Further, discovery is *not* limited to Meta's preferred fair use defense, and the Court should not allow Meta to limit Plaintiffs' ability to obtain discovery. The interrogatories in Set 3 go to the core of the issues in this case. Plaintiffs seek information concerning, for example: Meta's decision-making in choosing to use shadow datasets, which is relevant to infringement (Nos. 21-23, 34, 35, 40-46, 48); the financial benefits Meta has received or anticipates receiving from use and distribution of Llama, which is relevant to Plaintiffs' damages (Nos. 19, 20, 28, 29, 30, 32); and licenses that Meta sought or considered with respect to copyrighted works, which is relevant to both infringement and damages (Nos. 22, 24, 31).

Meta separately objects to these interrogatories on numerosity grounds. Rule 33(a) grants the Court discretion to increase the number of interrogatories under a "good cause" standard. Courts "will allow additional interrogatories," especially in complex cases, "as long as they are not 'unreasonable or unduly burdensome or expensive, given the needs of the case, the discovery already had in the case, the amount in controversy, and the importance of the issues at stake in the

litigation.'" *E.g.*, *Protective Optics v. Panoptx*, 2007 WL 963972, at *2 (N.D. Cal. Mar. 30, 2007) (quoting Fed. R. Civ. P. 26(b)(2)(iii)). With Set 3, Plaintiffs served 22 interrogatories in excess of the presumptive limit of 25 with well over two months remaining before the end of discovery. Plaintiffs repeatedly sought Meta's informal agreement to a commensurate increase in the number of interrogatories, something "reasonable parties should [do] . . . in a complex case," instead of "engag[ing] in motion practice." *Protective Optics*, 2007 WL 963972, at *2. Prior stipulated filings in this case put Meta on clear notice that more than 25 interrogatories would be served. In the most recent Joint Case Management Statement, Plaintiffs expressly stated that they would "require expansion of the limits on the number of . . . interrogatories." Dkt. No. 71 at 10.[12]

While Meta seeks to restrict the number of interrogatories all Plaintiffs can serve on Meta, Meta itself has served 28 interrogatories on ***each*** Plaintiff. In the Joint Case Management Statement, the parties "agree[d] that the applicable [discovery] limits shall apply to each side, as opposed to each party." Dkt. No. 71 at 10. Meta therefore has served well over 200 interrogatories even though its "side" was limited to 25 total. And even if Meta's interrogatory count is measured plaintiff-by-plaintiff, Meta still served too many—28 per plaintiff, not the maximum of 25. Tacitly acknowledging this, Meta's most recent interrogatories include a statement anticipating answers to the extra three questions "if the parties reach agreement on, and the Court orders" additional interrogatories. Accordingly, there can be no prejudice to Meta (or Plaintiffs) in ordering the parties to answer each other's previously-issued interrogatories.

Defendants' Position

Plaintiffs mischaracterize Meta's position (both past and present) regarding several interrogatories in order to manufacture another dispute.

I. Meta Has Already Agreed to Respond to Interrogatory No. 15

Following the parties' October 16 meet and confer, and to avoid motion practice, Meta expressly agreed to substantively respond to Interrogatory No. 15. Ex. 1 at 5-6. It will do so by November 19th. The issue is therefore moot. Plaintiffs' motion practice over something that Meta has repeatedly told Plaintiffs it would do is a waste of time for the parties and the Court.

To be clear, Meta did not previously decline to respond to Interrogatory No. 15 on the ground that it consists of multiple subparts. Meta declined to respond to the Interrogatory on the ground that, because several of Plaintiffs' prior interrogatories contained distinct subparts, Plaintiffs had exceeded the 25 interrogatory limit by no later than Interrogatory No. 13. Meta

[12] Additionally, Plaintiff Farnsworth, only consolidated into this case on October 18, had issued zero interrogatories before Plaintiffs' third set. Interrogatory Nos. 18-43 are therefore within the bounds of Rule 33 for Plaintiff Farnsworth.

Meta served 25 additional interrogatories on Plaintiff Farnsworth *the day he was added to the case*. This was in addition to the multiple hundreds of interrogatories that Meta cumulatively served on Plaintiffs. Plaintiff Farnsworth has the same right as Meta to serve and seek responses to interrogatories before the December 13, 2024 discovery cut-off. Indeed, in the normal course, consolidation does not "completely merg[e] the constituent cases into one, but instead [] enable[es] more efficient case management while preserving the distinct identities of the cases and *the rights of the separate parties* in them." *Hall v. Hall*, 584 U.S. 59, 60 (2018).

informed Plaintiffs ***in February*** that they had violated Fed. R. Civ. P. 33(a)(1), as Courts in this District hold that "interrogatory subparts are to be counted as one interrogatory ... if they are logically or factually subsumed within and necessarily related to the primary question." *Synopsys, Inc. v. ATopTech, Inc*, 319 F.R.D. 293, 294 (N.D. Cal. 2016) (quoting *Safeco of Am. v. Rawstron*, 181 F.R.D. 441, 445 (C.D. Cal. 1998)).[13] Nevertheless, in the spirit of moving discovery forward and reducing the disputes being brought to this Court, Meta agreed to supplement Interrogatory No. 15. This dispute is moot and should be denied.

II. Meta Has Already Agreed to Supplement its Response to Interrogatory No. 17

Regarding Interrogatory No. 17, here too, Plaintiffs are litigating an issue that has already been resolved. Again, after the October 16 meet and confer, Meta agreed that it would supplement its response to Interrogatory No. 17 to remove any perceived ambiguity or qualification regarding its response that it is not relying on the advice of counsel defense. Meta will supplement this interrogatory by November 19 . This issue is and has been moot.

III. Plaintiffs' Additional Interrogatories are Untimely and Unwarranted

Plaintiffs' demand that Meta respond to 23 new, additional interrogatories, which it served several weeks ago without leave of Court, should be denied. As background, on October 9, Plaintiffs requested Meta's consent to serve an additional 23 interrogatories and advised that "[a]bsent such agreement, Plaintiffs will raise this request with Judge Hixson." Before Meta could consider Plaintiffs' request, let alone respond to it, Plaintiffs served Interrogatories 19 through 48 the next day. And did so without raising the issue with this Court. Meta disagrees that Plaintiffs should have any additional interrogatories, but in an effort to avoid burdening the Court, on October 21, Meta proposed that the parties stipulate to an additional five interrogatories for a total of 30 interrogatories per side. Plaintiffs rejected this proposal and offered no alternative. Again, Plaintiff did not seek relief from the Court.

Plaintiffs now seek to compel Meta to respond to Interrogatories 26 through 48, but they had no arguable authority to serve them in the first place. Indeed, it is black letter law that a party "may not serve additional interrogatories without [an] agreement or leave of the Court." *Jovanovich v. Redden Marine Supply, Inc.*, 2011 WL 4459171, at *3 (W.D.Wash. Sept.26, 2011); *FormFactor, Inc v. Micro-Probe, Inc.*, No. C-10-03095 PJH JCS, 2012 WL 1575093, at *8 (N.D. Cal. May 3, 2012) (noting that party "had no authority" to serve additional interrogatories absent an agreement or leave of the court). Because Plaintiffs failed to seek leave to serve Interrogatories 26 through 48 prior to the October 18, 2024 deadline to serve additional discovery (ECF 238) their request to enlarge their number of interrogatories is also untimely.

Above, Plaintiffs point to their representation to the Court back in *January* (ECF 71) wherein they previewed that they might seek an expansion of the interrogatory limits, as notice to Meta that they would do so. Even if this did consist of notice, which it did not, it demonstrates

[13] To illustrate the overbreadth of Plaintiffs' interrogatories, take for example Interrogatory No. 7 (with discrete subparts (b)-(f) counting as at least four) or Interrogatory 13 (with discrete subparts (a)-(e) counting as at least three.

that Plaintiffs had months to seek an expansion of interrogatory limit , but did not bother to do so. They are now raising the issue 10 months later and just 5 weeks before the fact discovery cutoff.

Plaintiffs also attempt to use Plaintiff Farnsworth as an outlet for additional interrogatories, but even though he just joined the litigation, he is bound by the same rules the parties stipulated to in January, i.e., allotting interrogatories per side as opposed to per party.[14] ECF No. 71 at 9 ("The Parties agree that the applicable limits [in the FRCP] shall apply to each side, as opposed to each party.")

Plaintiffs' request to propound additional interrogatories beyond Interrogatory 25 should also be denied as unwarranted. "Leave to serve additional interrogatories may be granted to the extent consistent with Rule 26(b)(2)." Fed. R. Civ. P. 33(a)(1). However, courts must decline enlarging the interrogatory limit if it finds that (i) the discovery sought is unreasonably cumulative or can be obtained from a source that is more convenient, (ii) the party seeking discovery has had ample opportunity to obtain the information by discovery, or (iii) the proposed discovery is outside the scope permitted by Rule 26(b)(1). Fed. R. Civ. P. 26(b)(2)(C)(I)-(iii). "In practical terms, a party seeking leave to ... serve more Interrogatories than are contemplated by the Federal Rules ... must make a particularized showing of what the discovery is necessary." *Castaneda v. Burger King Corp.*, No. C 08-4262 WHA (JL), 2009 WL 4282596, at *1 (N.D. Cal. Nov. 25, 2009) (quoting *Archer Daniels Midland Co. v. Aon Risk Servs., Inc. of Minnesota*, 187 F.R.D. 578, 586 (D.Minn.1999)). *Accord. Roe v. Frito-Lay, Inc.*, No. 14-CV-00751-HSG(KAW), 2016 WL 1639774, at *3 (N.D. Cal. Apr. 26, 2016) ("Plaintiff bears the burden to show that the information she expects to obtain by propounding the additional interrogatories is proportional to the needs of the case, the parties' relative resources, and other Rule 26(b)(1) factors."). Plaintiffs have utterly failed to make such a showing. To the contrary, the additional interrogatories are emblematic of Plaintiffs' chronic 11th hour overreach, and only serve to demonstrate why their request should be denied.

The Court modestly enlarged the discovery period to permit Plaintiffs to obtain "what more it is that you need for adjudication of the fair use issue," 10/4/24 Tr. at 18, yet few of Plaintiffs' proposed additional interrogatories touch on that topic—let alone *any* relevant subject matter. Instead, they seek, by way of example, the identity of employees who communicated with any

[14] Plaintiffs suggest that Meta violated the parties' stipulation of 25 interrogatories per side by serving *identical* interrogatories on each Plaintiff individually, but this is nonsensical. Meta could have served one copy of "common" interrogatories on all Plaintiffs, which would have had the same effect and carried the same obligations, but chose not to do so to avoid confusion about who was responding to what. Farnsworth received the same interrogatories as the other 12 Plaintiffs. If Plaintiffs actually believed this violated the parties' agreement they presumably would have mentioned it in the many months since Meta first served discovery on January 9, 2024, but did not.

In addition, Plaintiffs point to Meta's service of ***3*** additional interrogatories as somehow justifying their demand for ***23*** additional interrogatories, but Meta explicitly indicated that its additional interrogatories were conditioned on the parties reaching a compromise on additional interrogatories. If the interrogatory limits are not expanded (and they should not be), Meta does not expect Plaintiffs to respond to those 3 additional interrogatories.

third parties concerning training data generally (Rog 25), the identity of individuals who assisted in discovery (Rog 26), research grants for the development of Llama models (Rog 33), a description of Meta's communications with Eleuther AI (Rog 34), the identities of those who had access to source code (Rog 36), policies for preventing disclosure of confidential information (Rog 37), locations where "Plaintiffs may access current Llama Model versions for inspection and analysis" (Rog 38), criteria for who receives access to the Llama Models (Rog 39), complaints received by Meta from other individuals complaining about Meta's use of the training data at issue in the complaint (Rog 42), the purpose of Meta's SRT tool, which is used to solicit and provide legal advice (Rog 44), the identities of individuals or departments with access to the SRT (Rog 45), identification of any "communications, messages, or data" related to Llama that were entered into SRT (Rog 46), a description of Meta's document collection efforts (Rog 47), and a description of the contents of materials that Meta produced to Plaintiffs and that are in Plaintiffs' possession (Rog 48). Among other issues, none of these interrogatories is relevant or satisfies the standards set forth above for interrogatories in excess of the default number under the Rules. *See Khan v. Payton*, No. 20-CV-03086-BLF (PR), 2024 WL 3680715, at *2 (N.D. Cal. July 16, 2024) (denying request because interrogatories were irrelevant or impermissibly compound).

Other interrogatories target subject matter that is exclusively the provenance of privilege (e.g., "Describe any legal review or analysis regarding the use of the Shadow Datasets to train Llama Models" – Rog 35; "Describe any analysis or assessment undertaken to identify or quantify potential legal risks associated with using the Shadow Datasets" – Rog 43), or seek information that is cumulative of Plaintiffs' document requests or for which Meta has previously represented that such information does not exist or is not kept in the ordinary course of business (e.g., "Describe the profits associated with Your distribution, use or provision of the Llama Models, including a detailed explanation of the calculation of the profits" – Rog 28; "Identify and describe all costs You incurred in the development, training, and distribution of the Llama Models" – Rog 30; see also Rogs 27, 29, 31-32). All of these Interrogatories are similarly contrary to Rule 26.

In short, even if the Court were to entertain Plaintiffs' untimely request to serve new discovery, Plaintiffs do not—and cannot—carry their burden to demonstrate why *these* additional interrogatories are needed.

Plaintiffs' Reply

Parts I and II: Plaintiffs appreciate Meta's willingness to supplement these interrogatories to moot the issues. However, this appears to be an about-face of Meta's prior position, and the Court should enforce Meta's promised relief.

Part III: Meta argues Plaintiffs "had no arguable authority" to even *serve* interrogatories additional to the presumptive 25-interrogatory limit, even though Meta acknowledges it has also served 28 interrogatories to every Plaintiff without the leave of court they say is required. Further, contrary to Meta's contention, Plaintiffs are not "raising the issue" of additional interrogatories "just 5 weeks before" the end of discovery. Plaintiffs served the interrogatories *on October 11*, with over two months to go before the cut-off. Indeed, Plaintiffs had no choice, given the time left in discovery, but to seek Meta's "informal agreement" to the excess interrogatories as something "reasonable parties should do" in a complex case. *Protective Optics*, 2007 WL 963972, at *2. Further, Meta's attempt to prove irrelevance of Plaintiffs' interrogatory topics demonstrates

their *relevance* to fair use and the other obviously essential issue at summary judgment--*proving Plaintiffs' own infringement claim*. Meta's attempt to case the interrogatories as invasive of privilege is also unavailing, as Meta admits in this briefing it still has not disclaimed an advice of counsel defense. Indeed, even if advice of counsel were not at issue, inquiries into Meta's identification of legal risks do not invade the privilege. *See, e.g.*, *Illumina Inc. v. BGI Genomics Co.*, 2021 WL 2662074, at *4 (N.D. Cal. June 29, 2021) (a "businessperson's statement that the company needs to assess legal risk as part of its SWOT analysis is not privileged," and a "slide deck that mentions potential legal risk twice in passing is not how a company asks for legal advice").

Defendant's Response to Plaintiffs' Reply

Parts I and II: There is no about-face. Meta expressly represented to Plaintiffs in an October 21 email that it would supplement responses to these interrogatories. *See* Ex. at 5-6 ("Meta intends to supplement its response to [Rog 15]" and "Meta will supplement its response [to Rog 17] in the manner requested"). There is simply no excuse for Plaintiffs wasting the parties' and the Court's time litigating issues that were resolved weeks ago.

Part III: Meta acknowledged above that Plaintiffs raised the issue *with Meta* on October 9. Nothing but Plaintiffs' own decision making caused them to (1) serve additional interrogatories beyond the limit before seeking leave of the Court, (2) reject Meta's offers of compromise, or (3) avoid seeking leave prior to the parties' stipulated deadline to serve additional discovery, as Plaintiffs indicated they would on October 9. Plaintiffs' demand to compel Meta to respond to additional interrogatories that Plaintiffs lacked authority to serve is untimely, and Plaintiffs cite no legal authority to the contrary.

Plaintiffs likewise fail to offer any justification for the additional interrogatories. They simply assert that "Meta's attempt to prove irrelevance [] of the interrogatory topics demonstrates their relevance" without any explanation. Plaintiffs have the burden of showing that the additional interrogatories are necessary and comport with Rule 26. *Roe*, 2016 WL 1639774, at *3. They cannot meet their burden with conclusory assertions of relevance, which are belied by the subject matter of the interrogatories. As noted above, most of the interrogatories concern facially irrelevant subject matter (indeed, much of it is "discovery on discovery"), topics that are cumulative of other discovery, or that seek to invade privilege.

As to Interrogatory Nos. 35 and 43, Plaintiffs insinuate that there is some ambiguity about whether Meta will rely on an advice of counsel defense, and that this entitles them to require Meta to describe "any legal review or analysis" and "any analysis or assessment undertaken to identify or quantify potential legal risks" concerning use of certain datasets. There is (and has been) no ambiguity: Meta has represented to Plaintiffs that it is *not* relying on the advice of counsel defense. Plaintiffs' reliance on *Illumina* is also misplaced. That decision addressed whether a marketing plan that contained "[a] businessperson's statement that the company needs to assess legal risk" was privileged. *Illumina*, 2021 WL 2662074, at *4. In holding that the document was not privileged, the court remarked, "[i]mportantly, the document does not contain any such assessment of legal risk" and did not reflect a solicitation of legal advice. *Id*. Here, legal "review," "analysis," and "assessments" are precisely what Plaintiffs seek.

ISSUE #4: ISSUES INVOLVING CONSTRUCTION OF PLAINTIFFS' RFPS

Plaintiffs' Position

I. The Court Should Order Meta To Produce All Documents Responsive To RFPs That Meta Has Narrowly or Otherwise Wrongly Misconstrued.

Meta refuses to produce documents responsive to several of Plaintiffs' Requests for Production ("RFPs") based on improper distinctions or misinterpretation of the RFPs. *See* Exs. __, __ (excerpts of Meta's responses to RFPs). Those RFPs seek documents relevant to four topics: (1) Meta's licensing agreements for training data (RFP Nos. 64 and 77); (2) the mechanisms or tools by which Meta filters its training data either to identify potential licensors of data or to exclude (or not) copyrighted data from its training datasets (RFP No. 45); (3) financial records or financial projections concerning Llama (RFP Nos. 46 and 53); and (4) certain features programmed into Llama's source code (Nos. 54 and 59). The scope of each RFP is proper as written and the Court should order Meta to produce all non-privileged documents responsive to them.

RFP No. 64 requests "Documents and Communications sufficient to show each instance within the last three years where [Meta] ha[s] licensed copyrighted works for Meta's commercial use." A related request, **RFP No. 77**, seeks communications "concerning any licensing [of] copyrighted works that were used to train the Meta Language Models." Meta objects to RFP 64 as unduly burdensome and not proportional to the needs of the case, and has committed only to "making a production regarding the licensing of copyrighted textual works for Meta's use in training the Meta Language Models."[15] Meta has also narrowly constructed RFP No. 77, asserting it is unduly burdensome because Meta cannot know which works are copyrighted, and Meta will therefore search only for "communications concerning Meta's negotiations of licenses for datasets, if any, that were used to train the Meta Language Models." Meta's construction of both RFPs is impermissibly narrow. Meta's stated intent to "mak[e] a production regarding" licensing of the models in response to No. 64 is insufficient, as Meta is clearly not promising a *comprehensive* production sufficient to show to the scope of its licenses. And searching for communications concerning licensing of copyrighted works should not be restricted to "negotiations" for "datasets," where Meta may have communications that never achieved the status of "negotiations" over licensing and Meta may have licensed copyrighted material that is not in a "dataset." Further, Meta has not attempted to show or explain why No. 77 is overly burdensome as to Plaintiffs' use of the term "*copyrighted* works"—to do so, Meta would have to show there is such a high volume of licenses in general that Meta cannot produce communications concerning *all* licenses.

Meta's limitation in producing only licenses related to "copyrighted textual works," and its communications related to "negotiations," prejudice plaintiffs' ability to demonstrate and calculate damages. The Court should require Meta to provide all documents responsive to the RFPs to

[15] Meta also attempts, without citation to authority, to cast Plaintiffs' request for a proper production in response to RFP No. 64 as untimely. Plaintiffs raised the issue in a deficiency letter to Meta on October 9, three business days after the Court extended the discovery deadline by 60 days. There is no waiver to challenging objections to RFPs in general, and especially not under the circumstances here. Even if a waiver could possibly apply, Plaintiffs issued the related/overlapping RFP No. 77 on August 29, 2024, also well within the discovery period. There is no prejudice to Meta in requiring responsive to both RFPs on similar topics.

provide Plaintiffs sufficient information to compare the nature and value of value of licenses Meta *has* paid for with potential licenses Meta could have negotiated in return for the right to copy the Infringed Works.

RFP No. 45 requests documents concerning "any licensing, accreditation, or attribution mechanism, or similar tool for crediting, compensating, or seeking consent from owners of copyrighted works that were used to train the Meta Language Models." Meta has refused to produce responsive documents on the basis that there is no such "tool," even though Plaintiffs explained that they are asking for documents related to any process, program, or method by which Meta sought to filter out copyrighted data from Llama training datasets. Meta has contradicted its own representation by stating there is a "tool used to visualize data" that, as Meta has explained it to Plaintiffs, helps Meta programmers accomplish the filtering Plaintiffs' request contemplates. In response, Plaintiffs offered as a compromise that Meta screenshot the "tool used to visualize data" that Meta described, but Meta still refused to do that. The methods, processes, or "tools" by which Meta identified and filtered (or did not filter) copyrighted data from datasets is clearly relevant to Plaintiffs' infringement claim, and Meta should be ordered to produce all documents responsive to RFP No. 45.

RFP Nos. 46 and 53 request "Documents and Communications sufficient to show [Meta's] actual or projected income from the sale or licensing of the Meta Language Models," and "Documents and Communications Concerning any income statement, balance sheet, or statement of cash flows, Concerning any of the Meta Language Models." Meta has produced, in total, only one document relevant to financial projections for Llama and a minimal set of other documents related to a budget for licensing costs and income from licensing Llama, despite the fact that the company has invested *billions* in AI technology. In the parties' meet and confer on October 16, Meta confirmed there were "financial-related documents" in its possession that it had withheld, but Meta will delay producing those documents in response to the set of RFPs Plaintiffs served on October 9, 2024. *See* Ex. C. Meta has not confirmed it provided all documents responsive to the RFP Nos. 46 and 53, which Plaintiffs served in February and March 2024. The Court should order Meta to fulfill its obligations to search for and produce all documents responsive to Nos. 46 and 53 immediately, not on Meta's chosen, dilatory timeline.

Finally, **RFP Nos. 54 and 59** request documents concerning, respectively, "any decision by [Meta] to not develop an interface for end users to interact with any of the Meta Language Models" and "the ability of any Meta Language Model to output fictional works." Meta objects to even *searching for* documents responsive to these requests about key features of Llama based on pure sophistry. Meta objects that Meta cannot search (and thus has not searched) for documents responsive to RFP No. 54 because Meta ultimately *has* decided to develop a Llama interface for end users. In the October 16 meet and confer, Plaintiffs pointed out that an ultimate decision to include a user interface does not mean there were not previous internal decisions *not* to include one as successive versions of Llama have been developed or released. Meta's decision-making about how Meta users or the public at large would interact with Llama is clearly relevant to Plaintiffs' claim for infringement. Similarly, Meta has objected to No. 59 on the ground that the term "fictional works" is unclear. Plaintiffs have explained that "fictional" has its traditional, dictionary-defined meaning, and "works" refers to copyrighted works. The terminology is not vague or unclear, and certainly not so vague or unclear as to justify Meta's failure to initiate a search. The Court should require Meta to search for and produce documents responsive to both RFPs.

II. Meta Has Applied an Overly Restrictive Relevant Time Period to its Discovery Responses.

In its responses and objections to all of Plaintiffs' RFPs (and its productions), Meta unilaterally and improperly limited the relevant period to January 1, 2022, to the present. As an attempted explanation, Meta's attorneys told Plaintiffs during their October 16 meet and confer that work on Llama started no earlier than 2022. Meta's artificial limitation on its searches is inappropriate not only because the proposed class period begins on July 7, 2020 (three years before Plaintiffs first filed their Complaint), but because any time at which Meta copied—or discussed copying— copyrighted material that was eventually used to train Llama is relevant to Plaintiffs' claim. Such copying or discussions may have occurred before 2022 even if development of Llama did not start until that year. Further, Meta has provided Plaintiffs no proof aside from counsel's naked assertion that 2022 was the year development started. If there are no responsive documents in earlier periods, then there is *de minimis* burden in simply running searches for any responsive documents.

The Court should require Meta to run searches (and produce documents from) as far back as necessary to capture all instances in which Meta copied—or discussed copying—copyrighted data that was used to train Llama, or, at a minimum, as far back as the beginning of the class period.

Defendants' Position

Plaintiffs' complaints about Meta's RFP responses are either meritless or moot. Plaintiffs' requested relief should be denied.

RFP No. 64. This Request, to which Meta responded on April 20, 2024, seeks "Documents and Communications sufficient to show each instance within the last three years where You have licensed copyrighted works for Meta's commercial use." Plaintiffs raised no issue with Meta's response to this Request for over five months, and thus Plaintiffs' complaints come too late. In any event, Meta has already told Plaintiffs that it will be producing documents regarding the licensing of copyrighted textual works for Meta's use in training the Meta Language Models (i.e. the Llama models at issue in this case) in response to RFPs served more recently by Plaintiffs. Meta's forthcoming production reflects the scope of potentially relevant materials in this case, and Plaintiffs' request should be denied.

RFP No. 64 is unreasonably overbroad, as it would require Meta to turn the company upside down and identify *every single instance* within the last three years where Meta has licensed *any* copyrighted work, of *any type* and for *any commercial purpose*. Setting aside Meta's Language Models,Meta's business involves a broad range of business interests that might license copyrighted material far outside the scope of this case. For example, if Meta licensed a song to use in an advertisement for the Meta Quest VR device or if it used code for an Instagram update under an open source license, that would arguably be responsive, yet would have no bearing on any issues in this case, e.g., whether Meta training its Llama models on allegedly copyrighted material is fair use. Proving the overbreadth of Plaintiffs' request, during the October 16 meet and confer, Meta asked Plaintiffs whether they are seeking within this request documents concerning, for example, Meta's licensing of a CLE video for presentation to its legal team. Plaintiffs said yes. Plaintiffs provide no explanation for how licenses for songs in commercials, open source source code, or CLE videos would bear on damages or any other issue in this case. Thus, this request as written is unduly burdensome and wildly disproportionate to the needs of the case, would require Meta to

canvass its entire business to root out every license of a copyrighted work for any purpose–an impractical, if not impossible task, and one with no identifiable utility to Plaintiffs. The lone claim in this case is about whether training an AI model using datasets that allegedly contain Plaintiffs' books is copyright infringement and if so, whether it is a fair use. There is no proportionality between this claim and the incredibly broad reach of and irrelevant materials sought by this RFP. Plaintiffs' motion on RFP No. 64 should be denied.

RFP No. 77. This Request asks for "Communications Concerning any licensing copyrighted works that were used to train the Meta Language Models." To the extent there is a plain reading of this grammatically unsound Request, it refers to communications concerning actual "licensing" of copyrighted works that "were used" to train the Llama models. With respect to copyrighted textual works (which are the subject of Plaintiffs' claims), no such licenses exist and thus there was nothing to produce. Accordingly, the motion to compel must be denied.

To the extent that Plaintiffs now seek to rewrite this Request to encompass communications concerning *prospective* licenses or communications concerning material that may have been *contemplated* for use but "were [not] used" to train the models, that is not what this Request seeks. Meta notes that Plaintiffs' new RFP No. 130 covers prospective licenses and, subject to Meta's forthcoming objections and responses, responsive material is in the process of being reviewed and produced in response to that RFP No. 130.

RFP No. 45. This request, to which Meta responded on February 23, 2024, seeks "All Documents and Communications Concerning any licensing, accreditation, or attribution mechanism, or similar tool for crediting, compensating, or seeking consent from owners of copyrighted works that were used to train the Meta Language Models." Meta has consistently told Plaintiffs that Meta has not developed any such mechanisms or tools. Plaintiffs misrepresent to the Court that "as Meta has explained it to Plaintiffs, [the visualization tool] helps Meta programmers accomplish the filtering the request contemplates." Meta made no such statement and tellingly none is cited. Because there is nothing to produce in response to this request, the motion to compel must be denied.

RFP Nos. 46 and 53. As to these Requests, Plaintiffs once again misrepresent Meta's statements during the parties' October 16 meet and confer. Meta never "confirmed there were 'financial-related' documents in its possession that it had withheld" that were responsive to these Requests. Meta is not withholding responsive documents for these two RFPs. Plaintiffs served a number of additional RFPs (not at issue here) on other financial related documents and Meta will be producing documents in response to those requests in due course. Plaintiffs requested relief should be denied.

RFP No. 54. This Request, which Meta responded to on April 20, 2024, seeks "All Documents and Communications Concerning any decision by You to not develop an interface for end users to interact with any of the Meta Language Models." This request presupposes that there was a decision by Meta *not to do something* (*i.e.*, not releasing a user interface for Meta's LLMs), but in fact Meta *did do something* (*i.e.,* released a user interface for Meta's LLMs). Setting aside the fact

that there cannot be responsive documents to the request as it is written, the requested documents (if any existed) would have no bearing on any issue in dispute and Plaintiffs have not demonstrated any. Instead, Plaintiffs try to rewrite the request to cover "decision making about how Meta users or the public at large would interact with Llama" and claim without explanation that would be "clearly relevant." Plaintiffs cannot rewrite their RFP seven months later. Plaintiffs' requested relief should be denied.

RFP No. 59. This Request, which Meta responded to on April 20, 2024, seeks "Documents and Communications Concerning the ability of any Meta Language Model to output fictional works." It is unclear what this vague Request seeks, and as construed by Plaintiffs this Request would be overbroad and unduly burdensome for Meta to comply with. In its broadest sense, anything that is not entirely true or factual is in some sense "fictional." Thus, if read with that understanding, Plaintiffs are seeking all documents and communications concerning the ability of the Llama models to produce outputs that are not factual or are untrue–subject matter that ranges from the propensity of LLMs to hallucinate (i.e., provide false information), to their ability to write stories or poems, none of which has any bearing on the issues in dispute here. Plaintiffs provide no limiting principle for this request. Plaintiffs' requested relief should be denied.

Time Period for Discovery Responses. Meta has consistently represented since its initial discovery responses over eight months ago that it would be collecting discovery responsive to Plaintiffs' discovery requests based on a time period from January 1, 2022 to the present. See, e.g, February 23, 2024 Responses to Plaintiffs' First Set of Requests for Production. Ex. . Meta did not begin development of the Llama large language models at issue in this case until the Fall of 2022, but nonetheless Meta took the expansive approach of producing documents going back to the beginning of 2022. Plaintiffs raised no concerns about Meta's time frame for eight months, only to suddenly demand in October (after the original fact discovery period would have closed) that Meta should be collecting discovery for a three year period prior to the complaint—a time period that was not even specified in Plaintiffs' discovery requests. See, e.g., December 27, 2023 Plaintiffs' First Set of Requests for Production, Ex. ("'Relevant Period' includes and encompasses all times relevant to the acts and failures to act which are relevant to the Complaint.")

As the Court admonished Plaintiffs last week, Plaintiffs should have raised any concerns about "date ranges . . a long time ago." ECF No. 252 at 2. To change the date range for document collection and production now would require Meta "redo" of all of Meta's document collection and production to date, including recollecting, ingesting into its discovery systems, and searching for documents in this expanded date range, which would take more time and resources in the short time period remaining in discovery. Plaintiffs' untimely and unnecessary request should be denied.

Plaintiffs' Reply

Meta refuses to comply with RFPs that encompass executed and prospective licenses for training data (*see* RFP Nos. 45, 64, 77). Meta generally says that it will "get to it later" in the context of RFP No. 130. But executed licenses for training data are responsive to RFP Nos. 64 and 77. Regarding RFP No. 77, Meta limits the scope of the request to licenses for textual works, but this is improper: licenses for other types of work would allow Plaintiffs' experts to evaluate the value of, and market for, training data, generally. Those datapoints may be especially helpful for expert analysis given the marketplace for training data is largely in its infancy.

Regarding RFP No. 45, Meta created a [redacted] This is plainly a "tool for . . . seeking consent from" copyright owners. Meta refuses to produce communications to the different content owners, even though those outreach efforts were documented within the tracker. The communications noted in the tracker clearly "Concern[]" that tool. Regarding RFP No. 59, Meta argues vagueness but admits it understands the scope of the request. The model's ability to write fictional stories is plainly relevant to factor four of the fair use analysis, as it goes to substitution in the marketplace for the copyrighted work.

Regarding date ranges, Meta embellishes the extent of its burden. No "re-do" is necessary—Meta never even searched for any documents that pre-date January 1, 2022 in the first place. Plaintiffs simply ask the Court to order Meta to confirm through discovery that responsive documents do not exist prior to January 1, 2022. Indeed, Meta/Facebook has been involved in the artificial intelligence race for a decade: https://www.fastcompany.com/3060570/facebooks-formula-for-winning-at-ai (June 2016 article discussing efforts related to AI and machine learning at Facebook and stating that AI "has become a vital part of scaling Facebook").

Defendant's Response to Plaintiffs' Reply

Regarding RFPs 45, 64, and 77, Plaintiffs cannot rewrite their requests for production to cover documents they wish they had more specifically asked for. Plaintiffs do not even attempt to defend their request for all licenses for copyrighted work in any business context at Meta in the past three years, because they cannot defend the overbreadth of their original request. On RFP 77, this case is about textual training data, not licensing of video, audio or other forms of data that might be used for training. The Court has already addressed Plaintiffs' demands for non-textual training data when it agreed with Meta's narrowing of 30(b)(6) Topic No. 8 in a discovery order last week. ECF No. 252 at 3.

Regarding RFP No. 45, Plaintiffs' response is to call a single document that Meta produced a "tool." [redacted] Plaintiffs identify nothing else that they claim is missing in response to RFP No. 45. Regarding RFP No. 59, Meta's argument illustrated the overbreadth and disproportionality of the potential scope of the request (which Plaintiff has never clearly defined). Plaintiffs for the first time raise "substitution in the marketplace", but have not developed any alleged connection to the fair use analysis. Finally on date ranges, Plaintiffs do not contest that they sat on Meta's consistent statements that Meta was objecting to documents that predated January 1, 2022 because the Llama models were not even in development until much later in 2022. If Plaintiffs believed they needed

a broader date range, Plaintiffs should have raised any concerns about "date ranges . . a long time ago." ECF No. 252 at 2.

ISSUE #5: SOURCE CODE AND TRAINING DATA

Plaintiffs' Position

One of the central issues in this case is the mechanism by which Meta trains its Llama models. No party disputes the relevance of this information. However, despite many attempts at remedial measures with Meta, certain components of this technical data—namely, the source code and the training data repositories—remain inadequately produced. Plaintiffs request an order requiring Meta to remedy the below issues.

I. Meta Must Provide Additional Llama Source Code.

Meta's production of Llama source code remains deficient notwithstanding Plaintiffs' repeated attempts to resolve this issue informally with Meta. For one, Meta still has not produced certain source code repositories, which Plaintiffs identified on October 4, 2024. These include: (1) Mitigation code (source code related to how Meta trains its models to identify copyrighted material to prevent its models from regurgitating that material); (2) Production code (code that customers actually use); and (3) Application code (code comprising a runnable computer program or system, including interface code and system infrastructure code). *See, e.g*, Dkt No. 206, Declaration of Jonathan Krein, at ¶¶ 6, 17.

Plaintiffs have strong reason to believe the missing code repositories exist, and that Meta just refuses to produce them. For example, Meta admits that it has developed "mitigations that tell the model not to respond to certain types of question." Chaya Nayak Dep. Tr. at 298:1-5. Yet, it has not produced any Mitigation code repositories. And, according to Plaintiffs' technical expert, "for the meta.ai. website to function, there must be APIs that facilitate access between the user interface and the production llama model(s) supporting the functionality available at that site." Krein Decl., Dkt. No. 206, ¶ 17. To date, Meta has not produced any Program code or Application code which would be required to support these core functionalities of Meta's products.

Further, based on Dr. Krein's additional review of the Llama source code since October 4, he has identified with additional specificity the categories of missing source code materials that Meta must make available. Dr. Krein lists these missing materials with specificity in his accompanying declaration. Ex. D.

Secondly, Meta refuses to produce the source code that integrates the Llama model with Meta's other offerings, a key initiative in Meta's efforts to commercialize its LLMs. *See* Meta_Kadrey_00046433 (". . . with Gen AI, we will supercharge our existing product offerings across Meta . . ."); Chaya Nayak Dep. Tr. at 289 (interpreting Meta_Kadrey_00046433 to mean "we intend to integrate generative AI into our existing offerings like Facebook."). Plaintiffs have served a request for production that would encompass this source code, but Meta refuses to make it available for inspection.

Meta should be ordered to produce or make available the aforementioned missing source code, as detailed in Dr. Krein's declaration.

II. Plaintiffs Are Entitled to More Fulsome Information About Llama's Training Data.

Plaintiffs have been seeking information about: (a) the training data for Llama models 1-3 (RFP Nos. 1-3); (b) documents and communications to/with/from a pirated books dataset called "LibGen" (RFP No. 7); and (c) a detailed description of *all* the data Meta has used to "train *or otherwise develop*" its Large Language Models (LLMs) (Interrogatory No. 1) (emphasis added).

The first Interrogatory sought information about, for example, how Meta obtained such data, and how much of it was used.

This is among the most relevant discovery in the case. The core claim is that each Plaintiff, and writers at large, were harmed when Meta stole – via scraping or downloading – copyright protected materials, including from known pirated sources, and used those materials *precisely for their expressive value* in 'training' or otherwise fine-tuning or operationalizing its models for use in Meta's suite of commercial products. To be sure, the parties have negotiated, and Plaintiffs have been able to obtain, important discovery in the data category. Were the only issue in this case the important binary question of whether Plaintiffs' copyrighted materials were in Meta's possession at a high level, the Court would have the answer from the data already produced. But it is also important that Plaintiffs be permitted discovery that goes to the *importance* of and breadth of use of the copyright protected material at issue, and that permits reasonable identification of the specific copyright protected materials used.[16] This '*in situ*' understanding and information relates to infringement, the fair use defense (for example, to the factors of the purpose of use, and the amount used), willfulness, and even ascertainability.

For these reasons, Plaintiffs do not believe it is enough for them to have access to the set of training data for Llamas 1-3, and submit they are entitled to information from Meta that identifies the iterations of copies of training data with copyrighted material or books within their possession, custody or control. Plaintiffs seek information on how many times Meta downloaded each copyrighted work, from where it downloaded each, when it downloaded each, and how it is using each copy.

Meta has not disputed that there are multiple locations that have iterations of training data, but raises a burden concern in light of its production of a set of training data for Llamas 1-3. To avoid any burden, Plaintiffs propose that Meta either (a) provide a declaration with this information (which Plaintiffs have proposed in the past) or (b) supplement Interrogatory 1 to provide this information.

Plaintiffs also had been amenable to receiving actual copies of data, but to narrow issues, to not dispute Meta's burden concern (while noting that the fact of burden suggests how important these materials are). Plaintiffs should be able to obtain relevant information about how copyright protected books are used by Meta, and a proxy set of information would involve discovery into the sources that contain the books.

Defendants' Position

Source Code. The Court should reject all of Plaintiffs' requested relief with respect to source code because it does not pertain to "existing written discovery." Dkt. No. 253. First, there are no existing–as opposed to *new*–RFPs served by Plaintiffs seeking source code. Second, and

[16] It is important to underscore that a books dataset may keep appearing and disappearing on places on the web or dark web, and to the extent Meta continues to pull updated information, this is relevant. It may not be possible for Plaintiffs to have a second-by-second update, *and Plaintiffs are not seeking duplicate copies of static materials*, but instead information about the various ways and locations Meta is using books corpuses or books that on which Meta's LLMs have been trained.

relatedly, the issues raised by Plaintiffs concern their more recent RFPs, and the parties have not yet engaged in any meet-and-confer process regarding the new issues raised by Plaintiffs here.

Plaintiffs' argument about source code tellingly does not identify a single RFP in their existing written discovery that seeks source code. This is because there are no such requests. To the contrary, source code is the subject of certain of Plaintiffs' Requests for Production served on October 9, 2024, which are not part of the "existing written discovery" and whose initial responses are not due until today. For example, RFPs 121 and 122 seek "All Documents and Communications, including source code relating to" both "production code for Llama models" and "application for the Llama models," mirroring the subject of Plaintiffs' premature arguments here. Ex. 4. Meta will be serving its initial responses to these overbroad requests today (which, as discussed below, seek an enormous amount of code and documents that are neither relevant nor proportional to the needs of the case). In any case, the parties have not even begun the meet-and-confer process on Plaintiffs' source code requests to understand what Plaintiffs actually need and whether the parties can reach agreement on an acceptable supplement to the existing source code production. Meta will work with Plaintiffs on those issues,[17] but Plaintiffs' attempt to seek premature adjudication on unripe issues by end-running the meet-and-confer rules is not the proper way to conduct discovery.

In May 2024, Meta did produce and make available in discovery source code repositories for Llama[18], not because they were responsive to existing discovery propounded by Plaintiffs (they were not) but to allow Meta (if needed) to later rely on those repositories to explain in more detail the mechanics of the training process. These repositories include: (1) a repository of source code [redacted]) for pre-training and post-training of the Llama models, (2) a repository of code and data for evaluating the performance of the Llama models, including code to run certain benchmarks, and (3) a repository of data processing recipes, including for text data, used for training Llama. Plaintiffs did not review that code until mid-August 2024 (and have since done separate inspections). After Plaintiffs' initial review of source code in mid-August, they sent a letter asking for additional information relating to issues and pull requests, which is textual information reflecting when individual engineers extracted source code files from the repositories. Despite this information having no relevance and not being responsive to any discovery requests served by Plaintiffs, Meta produced the issues and pull requests that it was able to locate for the produced code and informed Plaintiffs prior to their serving of this joint letter brief that it could not locate any purported "missing" issues and pull requests. Ex. 5 (11/5 Email Stameshkin to Simon.)

Plaintiffs have not requested a single printout of any of the source code they have reviewed, evidencing that their more recent source code reviews are directed at generating discovery disputes rather than gaining an understanding how the accused Llama training processes function. Indeed, the issues relating to allegedly missing materials, detailed in the 10-page declaration from Dr. Krein, were never disclosed to Meta prior to the evening of November 6 when Plaintiffs served the declaration with their portion of the present joint letter. (Krein Decl.) It is not clear if Dr.

[17] Earlier this week, ***prior*** to any request by Plaintiffs, Meta had already added one of the requested repositories. Dr. Krein's statement that there has not been any code added is false.

[18] Meta also makes certain of its Llama source code available to the public via open source licenses. That source code is equally accessible and downloadable by Plaintiffs.

Krein has any particular expertise in LLMs or if he or Plaintiffs actually understand the issues on which Plaintiffs are prematurely seeking relief. Plaintiffs claim for example that Meta did not produce "mitigations," but mitigations are a form of post-training or fine-tuning in which certain data is used to teach the model to respond in certain ways. Post-training and fine-tuning code was included in the above-mentioned [redacted] Meta is willing to work with Plaintiffs to ascertain with more specificity what they may be seeking beyond the post-training and fine-tuning code that was produced more than six months ago, and supplemented within the past month at Plaintiffs' request, but these issues should not be raised for the first time in a letter brief filed with the Court.

Nevertheless, to the extent the Court is inclined to rule on Plaintiffs' vague and prematurely-raised requests, the Court should reject them as not relevant and proportional to the needs of the case. The case alleges a single claim for copyright infringement based on allegations that Meta downloaded and used the Plaintiffs' copyrighted books as *inputs* to train the Llama models. Plaintiffs claim that, during the training process, the text of their works was ingested and thus copied to train the Llama models. (Dkt. 133, ¶¶ 28-29, 63.) Once the model is trained, it can produce natural language *output* responses in response to user requests. (*Id*.) But notably, the Court *dismissed* Plaintiffs' claims relating to the *output* of the Llama models. (Dkt. 56, at 1-3.) Thus, what happens *after* the Llama models are trained, *i.e.*, *after* the alleged acts of infringement have taken place, has no bearing on issues in this case. Plaintiffs thus do not explain how their requests for "production" or "application" code, which appear to pertain to source code for incorporating already-trained Llama models into other Meta products such as Facebook and Instagram, or "mitigations" that take place after the accused Llama training process occurs, are relevant or proportional to the needs of the case. The requests for production and application code, moreover, potentially implicate an enormous amount of source code that will be burdensome to collect and make available for inspection, while offering nothing probative on the sole remaining copyright claim relating to training using Plaintiffs' works.

Training Datasets. Plaintiffs' amorphous request for "more fulsome information" about training data should be denied. As the Court is aware, Plaintiffs are authors of various books who allege that their books were part of a dataset (commonly known in the AI research community known as "Books3"), which they allege was used as training data for Meta's Llama models. Meta produced that dataset in May 2024, and then supplemented its production to include further datasets that Plaintiffs claimed include additional books; Meta will continue to do so in the event new books-related datasets are used. In response to Plaintiffs' Interrogatory No. 1, Meta also provided a detailed supplemental response identifying approximately 70 datasets in addition to Books3, and for each dataset, the Llama model(s) with which it was used, the locations where Meta believes they were obtained, and a detailed explanation of the various considerations that go into Meta's selection and use of datasets. (Ex. 6 at 8:7-12:6.) Meta has complied with its discovery obligations with respect to datasets. Even Plaintiffs acknowledge above that they have obtained "important discovery in the data category."

Plaintiffs nonetheless demand that Meta should be required to "identif[y] the iterations of copies of training data with copyrighted material or books within their possession, custody or control." This request is overbroad, unreasonable and not proportional to the needs of the case for several reasons. First, Plaintiffs' request for "copyrighted material" appears to cover non-textual

training datasets such as digital images, music, and other non-book and non-text data not at issue in the case, a scope of discovery the Court already rejected when agreeing with Meta's compromise scope for 30(b)(6) topic no. 8. ECF No. 252 at 3. Second, there is no basis for Plaintiffs to demand that Meta produce multiple "iterations" and "copies" of training data. Meta has identified and produced copies of the *actual* book-related training datasets that allegedly include copyrighted works that were *actually* used to train the Llama models. Plaintiffs' request that Meta scour the entire company to determine if there are other stored and duplicative copies of those datasets – copies which would *not* have been the ones used to train Llama – is overly burdensome and not proportional to the needs of the case. Even on the issue of statutory damages – the only category of damages Plaintiffs could plausibly seek in this putative class action – Plaintiffs are entitled only to a single statutory damages under the Copyright Act for each work, regardless of the number of alleged infringements. 17 U.S.C. § 504(c)(1) (authorizing "*an* award of statutory damages *for all infringements involved in the action*, with respect to any one work, for which any one infringer is liable individually…") (emphasis added); *see also Louis Vuitton Malletier, S.A. v. Akanoc Solutions, Inc.*, 658 F.3d 936, 946 (9th Cir. 2011) ("With respect to copyright, 'when statutory damages are assessed against one defendant… each work infringed may form the basis of only one award, *regardless of the number of separate infringements of that work*.'") (citation omitted; emphasis added). The existence or number of duplicate and redundant copies of datasets, which were not used to train Llama, is not relevant to the issues in this case. Third, these datasets can be enormous in size, in some cases spanning multiple terabytes per dataset, which would place enormous burden on Meta disproportionate to the needs of the case to collect and produce multiple copies of datasets.

Finally, with respect to RFP No. 7, Meta is unaware of communications between Meta and a third-party website. To the extent Plaintiffs suggest that this RFP covers copies at Meta of datasets relating in some manner to this website, Plaintiffs are ignoring the plain language of the RFP.

Plaintiffs' Reply

On source code, whether Llama outputs copyrighted books in response to end users' requests is plainly relevant to Plaintiffs' "input" claims. First, Llama can only output copyrighted books if it has been trained on those copyrighted books. Thus, evidence of Llama's ability to output copyrighted material, and Meta's efforts to mitigate these infringing outputs, go straight to the merits of this case. Second, this evidence is relevant to Plaintiffs' claims that Meta does not just create datasets of copyrighted material for use as training data, but also that the model, itself, stores copies of this copyrighted material within it. Meta is also incorrect that its production and application code is irrelevant because there are no "output" claims in this case. This source code shows how Meta's entire generative AI program is a *commercial* endeavor—and the question of commerciality is obviously a critical inquiry under fair use. Plaintiffs are willing to engage in an ongoing discussion with Meta regarding the provided source code. In light of the "raise it or waive it" approach that Meta has taken in recent weeks, Plaintiffs needed to memorialize the ongoing issues with Meta's datasets, along with reserving rights if the issues persist.

On training data, Meta's responses ignore the specific points Plaintiffs made. Plaintiffs are not seeking information about videos or copyrighted materials that may have been used to train diffusion models, but instead information about how books were used in LLM training and operationalization, and how in turn the LLMs or book corpuses are used by Meta. Meta effectively concedes the relevance of this information for liability purposes, but argues that because copyright

damages are by work versus by copy, the information is irrelevant. This is incorrect. Plaintiffs are entitled to discover the range of copyright protected books that Meta ingested, including from sources that may be updated or that are in patterns of being taken down from the web. With reference to statutory damages, discovery on bad faith and willfulness is unquestionably relevant to where in the range of statutory damages Meta's conduct lies. Further, while Meta asserts potential burden in providing the sought information, Plaintiffs are not even seeking each copy of the books made, but rather a declaration or interrogatory answer with this information.

Defendant's Response to Plaintiffs' Reply

With respect to source code, Plaintiffs' reply arguments do not cite a single RFP that seeks Meta's source code–all but conceding that their arguments are not part of the "existing written discovery" to which the present motion is limited. Plaintiffs claim above that "Llama's ability to output copyrighted material" is relevant but they ignore the indisputable fact that Judge Chhabria almost a year ago **dismissed** their claims based on output of the Llama model. (Dkt. 56, at 1-3.) Plaintiffs further claim above that "the [Llama] model, itself, stores copies of this copyrighted material within it," but Judge Chhabria also dismissed this claim as "nonsensical" because "[t]here is no way to understand the LLaMA models themselves as a recasting or adaptation of any of the plaintiffs' books." (*Id*., at 1.) Because Plaintiffs have no good faith argument that their lone surviving copyright claim embraces output of the Llama models, their request for discovery on that topic should be denied. Plaintiffs also do not explain how source code has any probative value on the alleged "commerciality" of Llama. Meta's internal and highly technical source code is not probative on issues relating to commercialization of Llama.

With respect to training data, Meta appreciates that Plaintiffs have now clarified that they are not seeking discovery on non-book/non-text training data, despite the broader language of their earlier demands. Plaintiffs claim above to be seeking "information about how books were used in LLM training and operationalization, and how in turn the LLMs or book corpuses are used by Meta," but that is ***exactly*** what Meta has provided by producing the training datasets that allegedly contain all of Plaintiffs' books and that were actually used to train the accused Llama LLMs. Meta should not be ordered to scour its internal server infrastructure to locate and identify additional copies of training datasets which, even if they were to exist somewhere, were ***not*** the ones used for Llama training. Plaintiffs' unexplained assertion that they are relevant to "bad faith and willfulness" does not justify the substantial burden that would be imposed by the requested relief, sought by Plaintiffs for the first time near the end of fact discovery.